基 礎 統 計 学

第 5 回　分割表の分析

2017. 11. 10

2×2 分割表（四分表）

二つの項目 X, Y があり，それぞれ２つの分類

$X_1, X_2,\quad Y_1, Y_2$ をもつとする.

これらのデータは項目，分類ごとに集計して 2×2

分割表にまとめられる.

例

X (薬品）：新薬，偽薬
Y (効果）：あり，なし

2×2 分割表

		Y		
		Y_1	Y_2	計
X	X_1	n_{11}	n_{12}	$N_{1\cdot}$
	X_2	n_{21}	n_{22}	$N_{2\cdot}$
計		$N_{\cdot 1}$	$N_{\cdot 2}$	N

問題

X, Y には関連があるのかないのか？

独立性の検定

帰無仮説 (H_0) : X, Y には関連がない（独立である）
対立仮説 (H_1) : X, Y には関連がある（独立でない）

として検定を行う.

検定方法

- 表の各マスの数値が 5 以上のとき χ^2 検定
- 5 未満のマスがあるとき Fisher の正確確率検定

ピアソンのχ^2検定

各マスの期待度数（期待値）をE_{ij}とする.

Table: 期待度数

		Y		
		Y_1	Y_2	計
X	X_1	E_{11}	E_{12}	$N_{1\cdot}$
	X_2	E_{21}	E_{22}	$N_{2\cdot}$
計		$N_{\cdot 1}$	$N_{\cdot 2}$	N

χ^2 値の定義

X, Y には関連がない（独立である）とすると,

$E_{ij} = \dfrac{N_{i\cdot} \times N_{\cdot j}}{N}$, $i, j = 1, 2$ となる.

観測度数 n_{ij} と期待度数 E_{ij} との違いを次式で評価する.

χ^2 値

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

χ^2 分布

Fact

$n_{ij} \geq 5$ のとき，χ^2 は近似的に自由度 1 の χ^2 分布に従う.

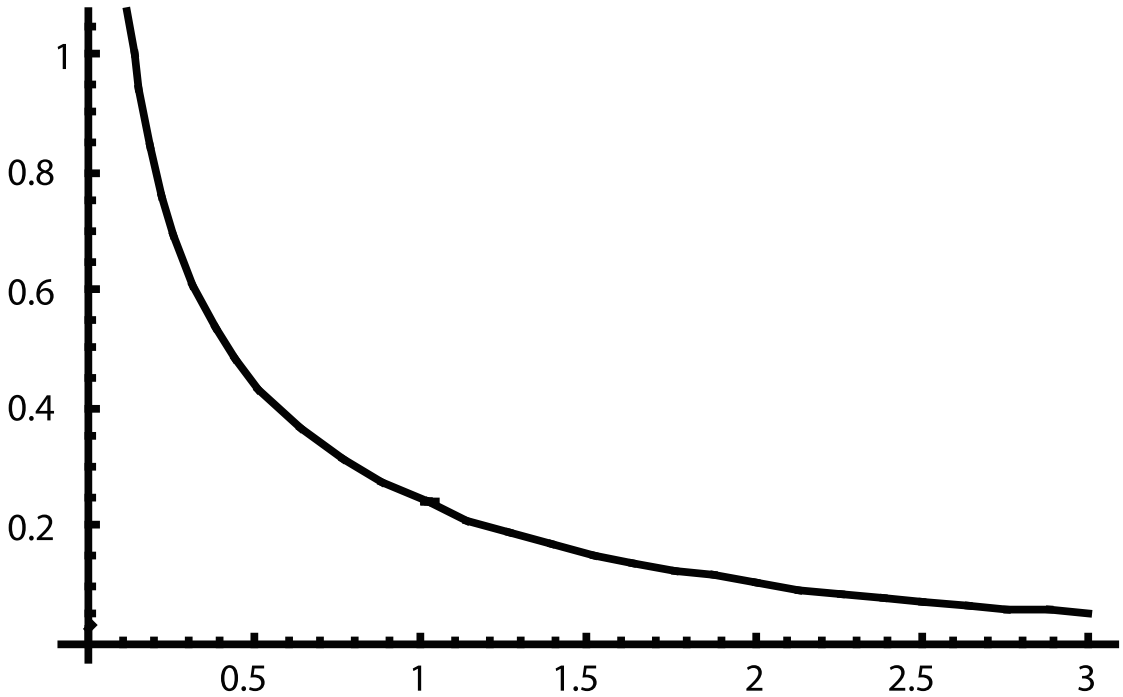

Figure: 自由度 1 の χ^2 分布

判定

- $\chi^2 > \chi_1^2(\alpha)$ のとき（p 値 $\leq \alpha$ のとき）

  (H_0) は棄却. したがって X, Y は独立でない（関連がある）.

- $\chi^2 \leq \chi_1^2(\alpha)$ のとき（p 値 $> \alpha$ のとき）

  (H_0) は棄却できない. 独立でないとは言えない.

$$\chi_1^2(0.05) = 3.84,\ \chi_1^2(0.01) = 6.64$$

R では p 値が出力される.

イェーツ補正

分割表のχ^2検定は近似的なもので本来の値から多少のずれがある.

そこで0.5の補正をして

イェーツ補正したχ^2値

$$\chi^2 = \sum_{i,j} \frac{(|n_{ij} - E_{ij}| - 0.5)^2}{E_{ij}}$$

を使うとより近似がよくなることが知られている.

この場合, 補正前よりp値は大きくなる.

補足

$k \times l$ の表でも独立性の検定ができる.

この場合は χ^2 値は自由度 $(k-1)(l-1)$ の χ^2 分布に従う.

EZR (R Commander) にて

1. medicine.csv を読み込む.
2. 統計解析 → 名義変数の解析 →

   分割表の作成と群間の比率の比較

カイ２乗検定をチェックしておくと，χ^2 検定が行われる.

Fisherの正確確率検定 (Exact Test)

あるマスの数が5未満のときは，χ^2 値の分布は χ^2 分布からのずれが大きくなり，正確な p 値にならない.

そこで，直接的に確率（p 値）を計算する方法が正確確率検定である.

注）一つのマスが5未満のときに Fisher's exact test は使われるが，この方法は5以上のときにも使える手法である.

確率の計算法(概略)

		Y		
		Y_1	Y_2	計
X	X_1	x	y	n_1
	X_2	z	w	n_2
計		m_1	m_2	N

計 n_1, n_2, m_1, m_2 が固定されている状況でデータの組合せで $x = a$ となる確率 $P(x = a)$ を計算すると

$$P(x = a) = \frac{n_1!n_2!m_1!m_2!}{N!a!b!c!d!}$$

となる. ここで $b = n_1 - a\,(= y)$, $c = m_1 - a\,(= z)$, $d = m_2 - b = m_2 - n_1 + a\,(= w)$.

判定

有意水準を α とし，$P(x < k) < \alpha/2$ となる最大の k と $P(x > l) < \alpha/2$ となる最小の l を求めると，
棄却域が $x < k$ または $x > l$ となる.

補足：全てのマス目が 5 以上の場合で Exact Test を使っても構わない.

再び **EZR** にて

EZR: 統計解析 → 名義変数の解析 →

分割表の作成と群間の比率の比較

から Exact test が行われる.

R: fisher.test(table)

リスク比とオッズ比

あるコホート (cohort) 研究（リスクファクターあるなしの群を追跡調査する前向きの研究）でつぎの表を得たとする.

		アウトカム		
		あり	なし	計
リスクファクター	あり	a	b	n_1
	なし	c	d	n_2
計		m_1	m_2	N

リスク比

母集団のリスク比は四分表のデータから推測できる.

リスクファクターありの場合の

アウトカムありの確率 $\dfrac{a}{a+b}$

リスクファクターなしの場合の

アウトカムありの確率 $\dfrac{c}{c+d}$

リスク比の推定値

$$RR = \frac{a/(a+b)}{c/(c+d)}$$

リスク比の解釈

- リスク比 RR はリスクファクターのあるなしで何倍の危険性があるかを示す数値
- $RR = 1$ ならリスクファクターのあるなしはアウトカムのありなしとは無関係 (独立) ということになる.

リスク比の信頼区間

母集団のリスク比は上記の RR で推定されるが,

その 95% 信頼区間はつぎで計算できる.

95% 信頼区間

$$\exp\left(\log RR \pm 1.96\sqrt{\frac{1}{a}+\frac{1}{b}-\frac{1}{a+c}-\frac{1}{b+c}}\right)$$

$\exp x = e^x$, $\log x =$ 自然対数.

- 1 が信頼区間からはずれていれば有意であると判定できる. （$RR = 1$ は棄却される.）

オッズ比（Odds Ratio）

以下は症例対照 (case-control) 研究の四分表とする.

		アウトカム		
		あり	なし	計
リスクファクター	あり	a	b	n_1
	なし	c	d	n_2
計		m_1	m_2	N

症例対照 (case-control) 研究（アウトカムの情報からリスクファクターを探る後ろ向きの研究）では前記のリスク比を計算することは不適切である. (何故か？)

オッズ

あるイベントが起こる確率を p としたとき，$\dfrac{p}{1-p}$ をオッズという．

オッズと確率の関係

$$\text{odds} = \frac{p}{1-p}, \qquad p = \frac{\text{odds}}{1+\text{odds}}$$

オッズは確率の別の表現と思える．

- $0 \leq p \leq 1$ だが，$0 \leq \text{odds} \leq \infty$ である．
- $\text{odds} = 1$ であれば起こる起こらないは半々ということを意味する．

リスクファクターありの場合のオッズ

$$\text{odds}_1 = p_1/(1-p_1)$$

とリスクファクターなしの場合のオッズ

$$\text{odds}_2 = p_2/(1-p_2)$$

の比

$$OR = \text{odds}_1/\text{odds}_2$$

を（母集団の）オッズ比という.

四分表のオッズ比

四分表のデータから母集団ぼオッズ比は推測できる．アウトカムありの群でのリスクファクターありのオッズは

$$\frac{a/(a+c)}{c/(a+c)} = \frac{a}{c}$$

アウトカムなしの群でのリスクファクターありのオッズは

$$\frac{b/(b+d)}{d/(b+d)} = \frac{b}{d}$$

オッズ比

Odds Ratio (OR) の推定値

$$OR = \frac{a/c}{b/d} = \frac{ad}{bc}$$

本来求めたいものは，リスクファクターによるアウトカムありなしのオッズ比であるが，これは上記のアウトカムによるリスクファクターありなしのオッズ比に一致することが数学的に証明できる.

OR は RR の代用にできる

本来知りたいのは RR であるが，case-control 研究では，直接 RR を推定できない．そこで OR を RR の代用とすることが多い．

定理

イベントの起こる確率 p_1, p_2 が十分小さいとき，

$$OR \fallingdotseq RR$$

証明の概略

p_1 = リスクファクターありの場合のイベントが起こる確率.

p_2 = リスクファクターなしの場合のイベントが起こる確率.

$p_i/(1-p_i) = p_1 + p_i^2 + \cdots \fallingdotseq p_i \ (i = 1, 2)$ より

$OR \fallingdotseq p_1/p_2 = RR$ □.

OR の 95% 信頼区間

OR の 95% 信頼区間はつぎで計算できる.

95% 信頼区間

$$\exp\left(\log OR \pm 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right)$$

この場合も 1 が信頼区間からはずれるとき有意となる.

（母集団のオッズ比は 1 でない.）

EZR にて

統計解析 → 名義変数の解析 →

分割表の作成と群間の比率の比較でオッズ比の信頼区間も計算される.